GDAŃSK UNIVERSITY
OF TECHNOLOGY

Subject card

Subject name and code	High Performance Machine Learning, PG_00050192		
Field of study	Informatics, Biomedical Engineering, Biomedical Engineering, Biomedical Engineering		
Date of commencement of studies	February 2023	Academic year of realisation of subject	2023/2024
Education level	second-cycle studies	Subject group	Optional subject group Subject group related to scientific research in the field of study
Mode of study	Full-time studies	Mode of delivery	at the university
Year of study	2	Language of instruction	Polish
Semester of study	3	ECTS credits	3.0
Learning profile	general academic profile	Assessment form	assessment

Conducting unit	Department of Computer Architecture -> Faculty of Electronics, Telecommunications and Informatics	
Name and surname of lecturer (lecturers)	Subject supervisor	dr inż. Tomasz Boiński
	Teachers	dr inż. Tomasz Boiński

Lesson types and methods of instruction	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	15.0	0.0	15.0	0.0	0.0	30
	E-learning hours included: 0.0						

Learning activity and number of study hours	Learning activity	Participation in didactic classes included in study plan	Participation in consultation hours	Self-study	SUM
	Number of study hours	30	6.0	39.0	75

Subject objectives	The aim of the course is presentation of methods for optimizing execution time of algorithms used in Machine Learning utilizing modern frameworks and hardware.

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K7_U04] can apply knowledge of programming methods and techniques as well as select and apply appropriate programming methods and tools in computer software development or programming devices or controllers using microprocessors or programmable elements or systems specific to the field of study, making assessment and critical analysis of the prepared software as well as a synthesis and creative interpretation of information presented with it	The student can use the Jupyter notebook environment for running and analyzing advanced computations in the field of machine learning. The student can implement advanced advanced training process handling mechanisms, as well as multi-process communication in the TensorFlow environment.	[SU4] Assessment of ability to use methods and tools [SU2] Assessment of ability to analyse information [SU1] Assessment of task fulfilment
[K7_W03] Knows and understands, to an increased extent, the construction and operating principles of components and systems related to the field of study, including theories, methods and complex relationships between them and selected specific issues - appropriate for the curriculum.	The student knows the architecture of GPU-equipped computing systems used for machine learning computations. The student can find bottlenecks among the consecutive stages of machine learning model training process.	[SW1] Assessment of factual knowledge
[K7_W42] Knows and understands, to an increased extent, the principles and trends in the analysis and design of local and distributed IT systems and the basics of computer modeling and computerization of complex cognitive and decision-making processes.	The student knows contemporary trends in design of computing systems dedicated for machine learning and can analyze their performance.	[SW1] Assessment of factual knowledge
[K7_U07] can apply advanced methods of process and function support, specific to the field of study	The student knows methods for reducing time of machine learning computations by choosing appropriate algorithms, vectorization, efficient utilization of available computing resources and parallelization.	[SU4] Assessment of ability to use methods and tools [SU1] Assessment of task fulfilment
[K7_U06] can analyse the operation of components, circuits and systems related to the field of study; measure their parameters; examine technical specifications; interpret obtained results and draw conclusions	The student can monitor the parameters and current utilization of the CPU, GPU, memory and hard drives with respect to specific processes in the GNU/Linux system. The student can profile performance of the individual operations in computational graphs used in machine learning.	[SU4] Assessment of ability to use methods and tools [SU2] Assessment of ability to analyse information [SU1] Assessment of task fulfilment

Subject contents

1. Introduction to the course, motivations for High Performance Computing in Machine Learning
2. Recap of primitives, loss functions and gradient methods used in Machine Learning
3. Methods for minimizing evaluation time of Machine Learning models
4. Methods for Machine Learning model training parallelization
5. Monitoring utilization of distributed computing resources used in Machine Learning
6. Techniques for profiling Machine Learning applications
7. Methods for distributed data representation and loading for artificial neural network training
8. Characteristics of hardware used for efficient Machine Learning
9. Parallelization capabilities of chosen Machine Learning frameworks
10. Case studies of artificial neural network training optimization in the fields of text analysis, visual and speech
recognition

Prerequisites
and co-requisites

Basic knowledge in the fields of parallel computing and machine learning, programming in Python.

Assessment methods
and criteria

Subject passing criteria	Passing threshold	Percentage of the final grade
laboratories	50.0%	50.0%
mid-term test	50.0%	50.0%

Recommended reading

Basic literature	B. Sjardin, L. Massaron, and A. Boschetti, Large scale machine learning with Python. 2016. M. R. Karim and Md. Mahedi Kaysar, Large Scale Machine Learning with Spark. Packt Publishing, 2016.

	Supplementary literature	F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, “On parallelizability of stochastic gradient descent for speech DNNs,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 235–239. J. Dean et al., “Large scale distributed deep networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1223–1231. J. Keuper and F. J. Preundt, “Distributed Training of Deep Neural Networks: Theoretical and Practical Limits of Parallel Scalability,” in 2016 2nd Workshop on Machine Learning in HPC Environments (MLHPC), 2016, pp. 19–26. Gupta, S.; Zhang, W. & Milthorpe, J. (2015), 'Model Accuracy and Runtime Tradeoff in Distributed Deep Learning.', CoRR abs/1509.04210.
	eResources addresses	Adresy na platformie eNauczanie:
Example issues/ example questions/ tasks being completed	Evaluating performance of chosen parallelization methods for artificial neural network training. Analyzing the influence of chosen optimization methods on model quality for a chosen application. Comparing capabilities of chosen Machine Learning frameworks based on a chosen application. Comparing performance of chosen hardware models for a chosen Machine Learning application.	
Work placement	Not applicable	